

Rec'd PCT/PTO 21 JAN 2005
10/521848

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization International Bureau



(43) International Publication Date
19 February 2004 (19.02.2004)

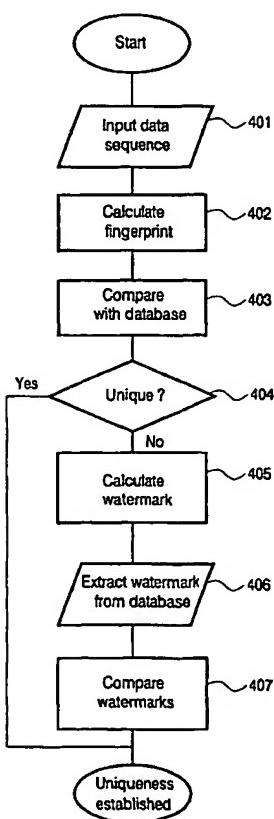
PCT

(10) International Publication Number
WO 2004/015629 A2

- (51) International Patent Classification⁷: **G06T 1/00**
- (21) International Application Number:
PCT/IB2003/002812
- (22) International Filing Date: 26 June 2003 (26.06.2003)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
02078079.7 26 July 2002 (26.07.2002) EP
- (71) Applicant (for all designated States except US): **KONINKLIJKE PHILIPS ELECTRONICS N.V. [NL/NL]; Groenewoudseweg 1, NL-5621 BA Eindhoven (NL).**
- (72) Inventors; and
- (75) Inventors/Applicants (for US only): **VAN DER VEEN,**
- Minne [NL/NL]; Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL). **BRUEKERS, Alphons, A., M., L.** [NL/NL]; Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL).
- (74) Agent: **GROENENDAAL, Antonius, W., M.; Philips Intellectual Property & Standards, Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL).**
- (81) Designated States (national): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.
- (84) Designated States (regional): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE,

[Continued on next page]

(54) Title: IDENTIFICATION OF DIGITAL DATA SEQUENCES



(57) Abstract: Identification of a digital media sequence is performed by an encoding and a decoding process. A sequence is received and its digital fingerprint is computed. A database lookup based on the fingerprint produces one or more matches that all resemble the computed fingerprint to a certain degree. If there is more than one match, at least one attempt is made to detect a watermark in the sequence. If a watermark is found, at least part of the watermark is extracted and used to select one of the matches among the sequences that resemble the media sequence to be identified.



ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declaration under Rule 4.17:

- *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii)) for the following designations AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW, ARIPO*

patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)

Published:

- *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

Identification of digital data sequences

The invention relates to a method and a system for enabling identification of a digital data sequence.

5 The handling of media content such as audio, images and image sequences have during the last decade or two entered the "digital era". More and more of the media content that is produced is produced, stored and transmitted via digital means such as computer storage media and digital transmission networks. Needless to say, this has lead to advantages as well as problems; in particular problems relating to legal issues such as proof
10 of ownership of the media content and the problem of unauthorized copying of the content.

Prior art includes at least two techniques to identify digital media content. These are watermarking and fingerprinting.

15 The watermarking technique can be summarized in that a unique identifier, i.e. a digital sequence of bits, is imperceptibly hidden in the content and can be extracted by a receiver for further processing, such as identification and authorization. However, a problem with the watermarking technique is that a large amount of bits needs to be embedded to allow globally unique identification, but it is very difficult to hide such a large identifier whilst making it impossible or very difficult to remove it from the media sequence in which it is embedded.

20 The fingerprinting technique involves recognizing unique features of a digital media sequence representing the content and converting these into a, ideally unique, bit sequence, i.e. a fingerprint. This fingerprint can be compared with other fingerprints and thereby identify the content in relation to other media sequences. However, a problem with fingerprinting is that a particular fingerprint might match two or more fingerprints of media
25 sequences. This problem is further accentuated when the fingerprinting technique involves ignoring "unreliable" bits in the fingerprint, i.e. when a certain level of robustness with respect to noise etc is needed.

In prior art, such as disclosed in UK patent application published with number 2 361 136, the watermarking and fingerprinting techniques have been combined in order to

improve identification of digital audio/video streams. In order to improve the procedure of proving provenance of a digital data stream, an identifying code in the form of a watermark is inserted into the data and a signature is also calculated based on the data. The watermark and the signature hence provides two independent means of proving provenance.

5

The object of the present invention is to provide a solution to the problem of how to simplify identification of digital media sequences.

10 The object is achieved according to two aspects by way of methods systems and computer programs according to the appended claims.

In some detail, there is provided according to a first aspect of the invention, a method, a system and a computer program for identifying a first digital data sequence. The method comprises calculating a first digital fingerprint based on at least part of the first sequence. This fingerprint is then compared with at least a second fingerprint, which is 15 associated with at least another, second digital data sequence. Depending on a result of the comparison, at least one digital watermark associated with the respective first and second data sequences is compared and, resulting from the comparison it is thereby possible to establish an identity of the first data sequence.

According to a second aspect of the invention there is provided a method, a 20 system and a computer program for enabling identification of a first digital data sequence. The method comprises calculating a first digital fingerprint based on at least part of the first sequence. This fingerprint is then compared with at least a second fingerprint, which is associated with at least another, second digital data sequence. Depending on a result of the 25 comparison, the watermark associated with the first sequence is stored for further use in enabling the identification of the data sequence.

Moreover, the use of the watermark may involve using watermark information that is calculated in dependence of the information contained in the first fingerprint or the difference between the fingerprint and fingerprints already stored in the database.

The technical effect obtained by the invention is hence that of enabling 30 identification of a data sequence by a conditional combination of watermarking and finger-printing, which can be seen as a hybrid identification method and system or, as the two aspects of the invention illustrate, in an encoding aspect and a decoding aspect.

When a content item, i.e. a digital sequence representing a media item or a part of a media item, is received for identification, a fingerprint is computed and added to a

database, preferably also together with appropriate metadata. The newly calculated fingerprint is compared with fingerprints that already exist in the database. If it is found that there is a sufficiently small distance between the newly computed fingerprint and an existing fingerprint, a watermark is embedded in the content of the data sequence. This watermark 5 preferably contains additional identifying information. This identifying information is then preferably also added as metadata to the database entry for that content item.

Identification of the media sequence can then proceed as follows. A sequence that is to be identified is received and its fingerprint is computed. A database lookup based on the fingerprint produces one or more matches that all resemble the computed fingerprint to a 10 certain degree. If there is more than one match, at least one attempt is made to detect a watermark in the sequence. If a watermark is found, at least part of the watermark is extracted and used to select one of the matches among the sequences that resemble the media sequence to be identified.

The watermark, or a part thereof, is then an identifier of the media sequence. 15 The identifier preferably represents the content item itself, but can also represent, e.g., the content owner for broadcast monitoring or otherwise provide an association between the media sequence and its provider or owner etc.

In fact, the invention may be divided into three separate sub-processes: an embedding process, an data base storage process and a detection (i.e. identification) process. 20 During the embedding process, the database is generated as described containing fingerprints and watermarks. One or more of the parameters that are contained in the information of the watermark is, in whole or in part, determined by the results from a comparison operation in which a fingerprint is compared with existing fingerprints in the database. In the database storage process, information of the watermark is appended. Example of such information is 25 type of watermark, watermark key, payload, etc. The storing of information in the database can be considered as a "training" process, in the sense that the information in the database will be more and more of use during later consultations of the database during future detection/identification processes.

The detection process is most simply described as an identification process 30 where a digital signal is identified using the database of fingerprints and metadata as well as the watermarks.

An advantage of the invention is that only a part of all considered content items need to be provided with a watermark. Only if there is a risk of a "clash" between two entries in the database, i.e. if there is a risk of confusing the media sequence with other media

sequences. This means that the total number of watermarked content items is lower than in a pure watermark-based identification system. As a result, the identifier in the form of a watermark to be embedded can be smaller, when compared with prior art, because it needs only be unique amongst the small number of content items that are watermarked. This
5 reduces the required capacity of the watermark.

The invention will now be described by way of preferred embodiments, with reference to a number of figures, where:

- 10 Figure 1 shows schematically a system according to the invention;
Figure 2 shows schematically a database structure in accordance with the invention; and

Figure 3 and 4 show a flow chart of a method according to the invention.

15 A method and a system which combines watermarking and fingerprint technology will now be described in some detail. As the person skilled in the art will recognize, the method and the system both involve processing means and memory units as well as communication means that are of a general character or of a more specialized
20 character. That is, general purpose computers with peripheral units such as hard disks, CD/DVD-recorders and connected to a digital network such as the internet may be utilized in an implementation of the invention. Specifically designed systems, comprising processors, memory units and communication means that are capable of only implementing the present invention are also feasible and are feasible to the person skilled in the art of designing
25 hardware and software in computing systems.

Figure 1 shows a schematic hardware view of a computing system 100 comprising a processor 101, a memory unit 102 and an input/output unit 103 that are interconnected via a bus 104. The system 100 is in connection with a digital communication network 105 through which information in the form of, e.g. digital media sequences
30 including audio, video or any other sequence that the system 100, a provider 106 and a user 107 wish to communicate. As the person skilled in the art will realize, the system 100 may include a number of additional units.

Turning now to a discussion of a method according to the invention, where a digital media sequence is to be handled by the system, the initial state of the system 100 will be defined.

Referring first to figure 2, illustrates a previously established database 200, which preferably is realized in the memory unit 102 of the system 100. The database 200 comprises information in the form of fingerprints 202 of digital media sequences as referenced by sequential numbers 201. The fingerprints 202 in the database 200 are, as the skilled person realizes a sequence of digits that have been calculated on the basis of the content of the respective media sequence. Linked to the fingerprints 202 are respective watermarks 203. However, not all fingerprints 202 have associated watermarks 203, as indicated by empty watermark positions 204 and 205, which illustrates the advantage of the invention, as presented above, that only part of all considered media sequences need to be provided with a watermark. Additional information, i.e. media content "metadata", associated with the respective media sequence, can also be accommodated in the database 200.

Continuing with the discussion regarding a method according to the invention, references will now be made to both figures 1, 2 and 3. Figure 3 shows a flow chart comprising steps performed by the system 100.

In an input step 301, a digital media sequence is input from the media sequence provider 106. In a following calculation step 302, a fingerprint is calculated. The calculated fingerprint, denoted by H_X , is in a comparison step 303 compared with fingerprints already present in the database 200, denoted by $H_{1\dots N}$ where $1\dots N$ denote fingerprints numbering between 1 and N.

In a decision step 304, it is decided, if the mathematical distance between the calculated fingerprint H_X and the existing ones $H_{1\dots N}$ is sufficiently large, i.e. if $M(H_X, H_{1\dots N}) > D_1$, where M defines a mathematical distance measure and D_1 is a limiting distance, then the fingerprint is defined as being unique. Then the process continues with a storage step 307 where the fingerprint is stored in the database and associated with the media sequence. That is, in the case of uniqueness of the fingerprint, recognition based on only fingerprints is successful.

However, if a possible non-uniqueness occurs, i.e. if $M(H_X, H_{1\dots N}) < D_1$, a watermark W_X is calculated in a calculation step 305 and embedded in the media sequence X in an embedding step 306. This watermark may contain additional identification information based on results obtained during the comparison step 303, i.e. a set of watermarks, which were used in the embedding of the corresponding multi-media signals. Based on this set of

watermarks, a new watermark is chosen. For example by choosing a new key or new payload for the watermark. This is then used for embedding in the new multimedia signal.

As for the case where the uniqueness was decided, in the decision step 304, the new fingerprint and associated watermark are appended to the database 200.

5 In an identification process, which may be performed by the user 107 when asking the system 100 for identification of a media sequence, the following steps may be performed by the system, as illustrated in the flow chart of figure 4.

In an input step 401, a digital media sequence is input to the system 100. In a following calculation step 402, a fingerprint is calculated. The calculated fingerprint, denoted 10 by H_x , is in a comparison step 403 compared with fingerprints already present in the database 200, denoted by $H_{1\dots N}$ where $1\dots N$ denote fingerprints numbering between 1 and N.

In a decision step 404, it is decided, if the mathematical distance between the calculated fingerprint H_x and the existing ones $H_{1\dots N}$ is sufficiently large, i.e. if $M(H_x, H_{1\dots N}) > D_2$, where M defines a mathematical distance measure and D_2 is a limiting 15 distance, then the uniqueness of the fingerprint has been established, i.e. the identity recognition has been based on fingerprints only.

However, if a possible non-uniqueness occurs, i.e. if $M(H_x, H_{1\dots N}) < D_2$, a watermark W_x is calculated in a calculation step 405. Watermarks 203 in the database 200 that are associated with the fingerprints 202 that were found to be mathematically close to the 20 fingerprint of the media sequence are then extracted from the database 200 in an extraction step 406. Finally, the calculated watermark is compared, in a comparison step 407, with these extracted watermarks and thereby establishing the uniqueness of the media sequence.

It is to be noted that, although the embodiments above discuss sequences of media data in a very general manner, it is understood that any type of media is relevant, and 25 can be exemplified by digital audio or video sequences as well as other sequences of data that is to be identified and/or associated with, e.g., an owner or provider. Any such sequence is considered to be equivalents and are within the scope of the appended claims.

Hence, to summarize, identification of a digital media sequence is performed by an encoding and a decoding process. A sequence is received and its digital fingerprint is 30 computed. A database lookup based on the fingerprint produces one or more matches that all resemble the computed fingerprint to a certain degree. If there is more than one match, at least one attempt is made to detect a watermark in the sequence. If a watermark is found, at least part of the watermark is extracted and used to select one of the matches among the sequences that resemble the media sequence to be identified.

CLAIMS:

1. A method for identifying a first digital data sequence, comprising:
 - calculating a first digital fingerprint based on at least part of the first sequence,
 - comparing the first fingerprint with at least a second fingerprint associated with at least a second digital data sequence,
- 5 – depending on a result of the comparison, comparing at least one digital watermark associated with the respective first and second data sequences and thereby establishing an identity of the first data sequence.
2. A method according to claim 1, further comprising:
 - 10 – calculating the at least one digital watermark, where the calculation is dependent on information contained in the first fingerprint.
3. A method according to claim 1, further comprising:
 - 15 – calculating the at least one digital watermark, where the calculation is dependent on information resulting from the comparison between the first fingerprint and the second fingerprint.
4. A system for identifying a first digital data sequence, comprising means for:
 - calculating a first digital fingerprint based on at least part of the first sequence,
 - 20 – comparing the first fingerprint with at least a second fingerprint associated with at least a second digital data sequence,
 - depending on a result of the comparison, comparing at least one digital watermark associated with the respective first and second data sequences and thereby establishing an identity of the first data sequence.
- 25 5. A system according to claim 4, further comprising means for:
 - calculating the at least one digital watermark, where the calculation is dependent on information contained in the first fingerprint.

6. A system according to claim 4, further comprising means for:

- calculating the at least one digital watermark, where the calculation is dependent on information resulting from the comparison between the first fingerprint and the second fingerprint.

5

7. A computer program including software instructions for controlling a computer to perform a method according to any one of claims 1-3.

8. A method for enabling identification of a first digital data sequence, 10 comprising:

- calculating a first digital fingerprint based on at least part of the first sequence,
- comparing the first fingerprint with at least a second fingerprint associated with at least a second digital data sequence,
- depending on a result of the comparison, storing at least one digital watermark associated with the first data sequence, thereby providing information enabling identification of the first data sequence.

15

9. A method according to claim 8, further comprising:

- calculating the at least one digital watermark, where the calculation is dependent on information contained in the first fingerprint.

20

10. A method according to claim 8, further comprising:

- calculating the at least one digital watermark, where the calculation is dependent on information resulting from the comparison between the first fingerprint and the second fingerprint.

25

11. A system for enabling identification of a first digital data sequence, comprising means for:

- calculating a first digital fingerprint based on at least part of the first sequence,
- comparing the first fingerprint with at least a second fingerprint associated with at least a second digital data sequence,

30

- depending on a result of the comparison, storing at least one digital watermark associated with the first data sequence, thereby providing information enabling identification of the first data sequence.

5 12. A system according to claim 11, further comprising means for:

- calculating the at least one digital watermark, where the calculation is dependent on information contained in the first fingerprint.

13. A system according to claim 11, further comprising means for:

- 10
- calculating the at least one digital watermark, where the calculation is dependent on information resulting from the comparison between the first fingerprint and the second fingerprint.

14. A computer program including software instructions for controlling a
15 computer to perform a method according to any one of claims 8-10.

1/3

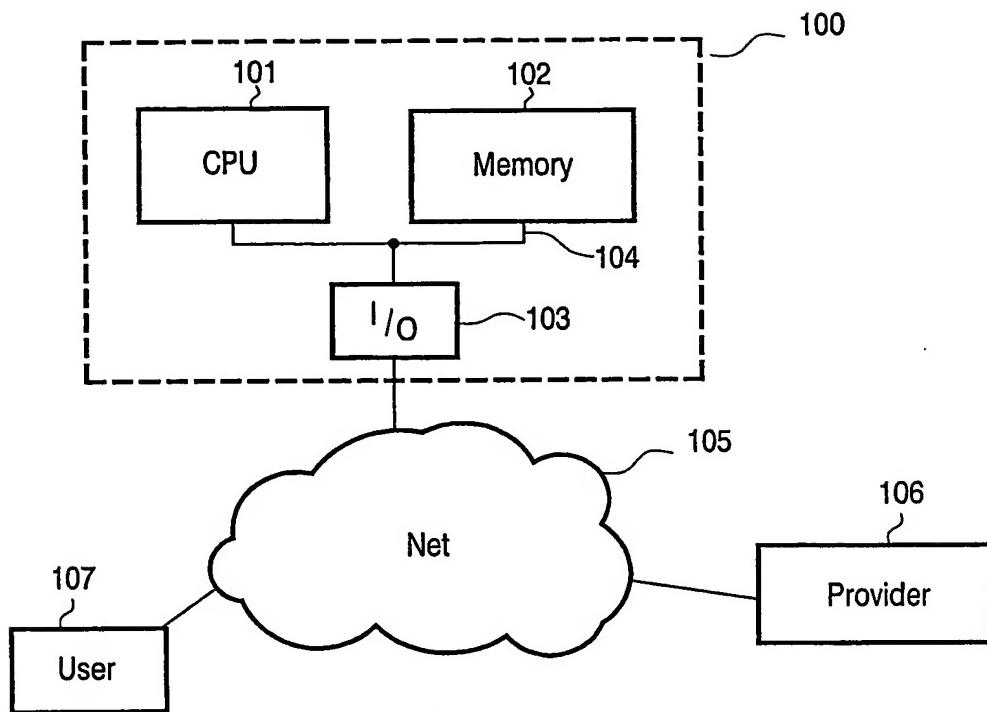


FIG. 1

The diagram shows a table with three columns: Row Number, Fingerprint, and Watermark. The table has rows numbered 1 through n. The first four rows are explicitly shown:

1	Fingerprint 1	Watermark 1
2	Fingerprint 2	<Empty>
3	Fingerprint 3	<Empty>
4	Fingerprint 4	Watermark 1
n	Fingerprint n	Watermark n

Arrows labeled 201, 202, and 203 point to the first three columns of the table. Labels 204 and 205 point to the right edge of the table. A dashed line labeled 200 points to the bottom edge of the table.

FIG. 2

2/3

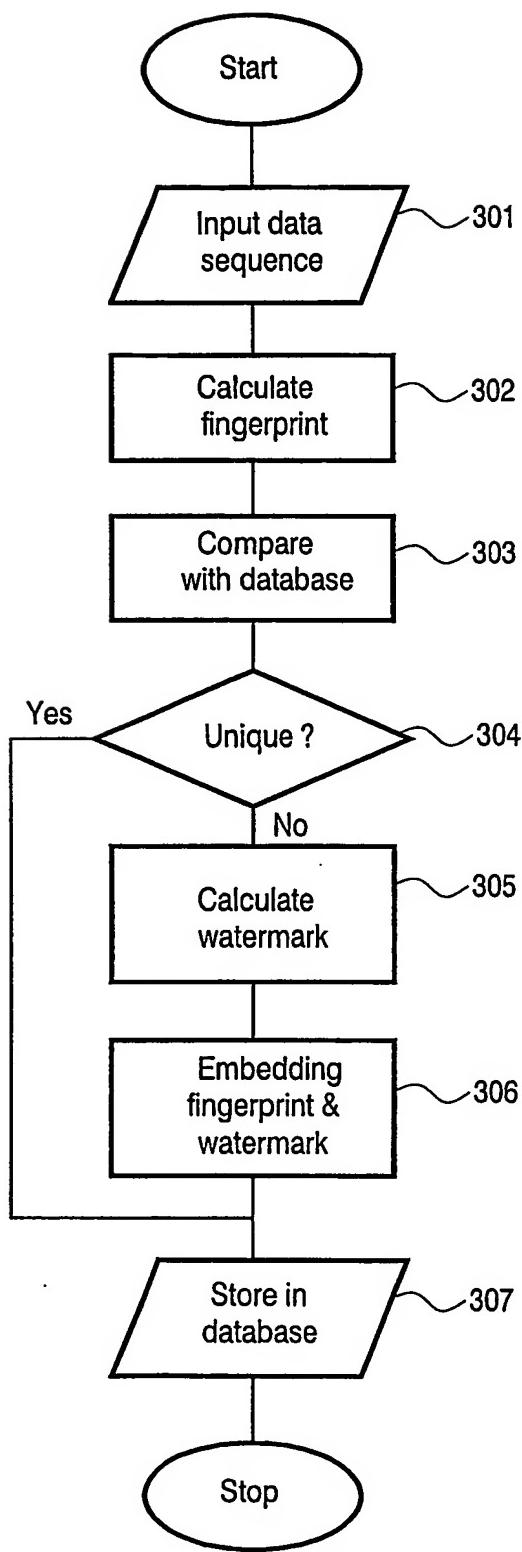


FIG. 3

3/3

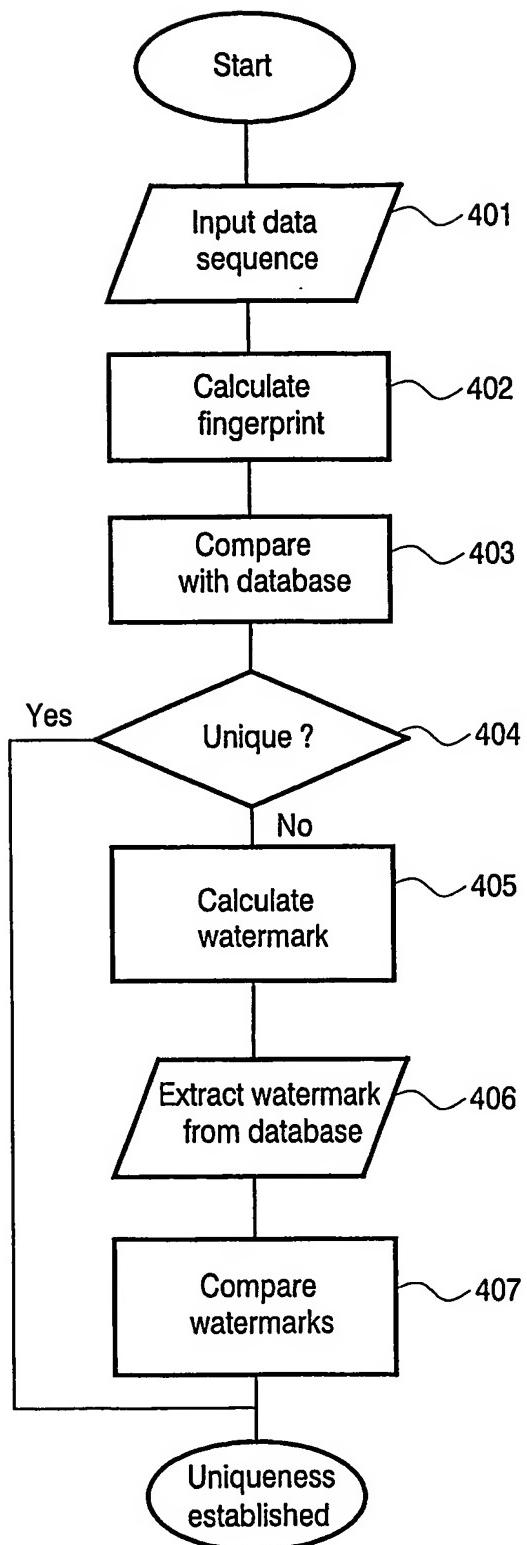


FIG. 4